Analysis of the expression profile of genes regulated by treatment with glucocorticoids and retinoic acid in lung cancer

Teresa Dalle Nogare[1]

Abstract

Different therapies are available for the treatment of lung cancer, among which the use of glucocorticoids (GC) and retinoic acid (RA). However, their role in the treatment of this specific cancer type is not well established. The aim of this project was to identify differences in expression patterns in both small lung cell cancer (SCLC) and non-small lung cell cancer (NSCLC) when subject to treatment with GC and RA, both alone and in combination with Suberoylanilide hydroxamic acid and azacitidine (AZA/SAHA). With this purpose, a subset of the public dataset (GEO accession GSE66245) was used for the analysis, focusing on data derived from a MYC-amplified cell line in which the BRG1 gene was genetically inactivated. Classification was performed using four supervised learning methods -Random Forest, LDA, LASSO and SCUDO- and the comparison of performance, quantified through the accuracy, suggested that Random Forest was the best performing algorithm. Functional analysis was performed with g:Profiler on the 100 most relevant genes by importance value as computed by the Random Forest. The higher ranked KEGG pathway was the retinol metabolism hsa00830 (p $=2.026 \times 10^{-2}$). The most significant GO term about the molecular function was retinoic acid 4-hydroxylase activity GO:0008401(p $=3.706 \times 10^{-3}$). Negative regulation of retinoic acid receptor signaling pathway GO:0048387 (p $=7.417 \times 10^{-3}$) was the most significant GO term concerning the biological process. Additional biological insights were gained from network-based analysis carried out using PathfindR, identifying an up-regulation of Cytochrome P450 26A1 and B1 (CYP26A1 and B1) in cells subject to treatment compared to the non-treated ones. The CYP26-mediated destruction of retinoic acid, suggested by the obtained results, provides a possible explanation for the limited clinical efficiency of RA in the treatment of many solid types of tumor, as previously reported (V. Hunsu et al., Int. J. Mol. Sci. 2021).

[1] *Master's degree in Physics, University of Trento*

Contents

Introduction

Lung cancer is a type of cancer that forms in tissues of the lung that can also spread to other parts of the body in the form of metastasis. Cells belonging to the primary tumor can break away from the region where they begin and travel through the lymph system or blood, reaching other parts of the body. It includes two major types of cancer:

- *small cell lung cancer* (SCLC) : is an aggressive and fast-growing cancer that forms in tissues of the lung. It represents about 10%-15% of lung cancer and is named after the characteristic oval shape of cancer cells observed under a microscope.
- *non-small cell lung cancer* (NSCLC) : is named after the shape of groups of cancer cells observed under a microscope. It represents about 80%-85% of lung cancer and includes adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.

Among the risk factors for both types of lung cancer, smoke and exposure to secondhand smoke are the most relevant ones while the major symptoms of this disease are chest discomfort and pain, cough and trouble breathing.

Microarray technology was used to detect both the expression and methylation profile of genes by collecting genetic material from samples of both SCLC and NSCLC cancer types. In this specific study, the purpose was to identify differences in expression patterns of SCLC and NSCLC after treatment with glucocorticoids (GC) and retinoic acid (RA) alone and in combination with Suberoylanilide hydroxamic acid and azacitidine (AZA/SAHA) from non-treated cells. The main peculiarities of FBS and the studied treatment types are:

- *Fetal Bovine Serum* (FBS) : is the liquid fraction of clotted blood from fetal calves, depleted of cells, fibrin and clotting factors, but containing a large number of nutritional and macromolecular factors essential for cell growth. They represent the reference samples.
- *Glucocorticoids and Retinoic acid* (GC-RA) : glucocorticoids are steroid hormons often used to treat cancer. Retinoic acid is a metabolite of vitamin A_1 that mediates the functions of vitamin A_1 required for growth and development. All-trans-retinoic acid (ATRA) is the major occurring retinoic acid which can be employed to treat cancer.
- *GC-RA combined with* AZA/SAHA : AZA/SAHA represents an amino group inserted into the hexyl linker moiety of the approved drug Vorinostat (SAHA) [1]. Vorinostat is a member of a larger class of compounds that can be used for cancer treatment.

Dataset preparation

In this work, studies were carried out on the dataset GSE66245 which is a super series composed of three sub-series:

- GPL13534 : methylation profile
- GPL17077 : expression profile
- GPL21185 : expression profile

Methylation profile was acquired exploiting genome tiling array making use of Illumina HumanMethylation450 BeadChip. For what concerns expression profiles, Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray was exploited.

For the purposes of this work, only a subset of data belonging to the original super-series was analysed. The suitable dataset, that involved expression profiles of both platforms GPL17077 and GPL21185, was obtained through a pre-processing of the whole super series. Concerning data belonging to platform GPL17077:

- The last 9 samples had an offset compared to the others that could not be discarded in the normalization procedure. Since data were probably obtained from another laboratory, and the source of error was considered to be due to the measurement technique, these data were discarded.
- The remaining 55 samples were acquired using two different cell line types, namely MYC amplification and BRG1-mutant. Since the purpose of this project was not to analyse the relationship between treatment and cell line type, rather it aimed at identifying genes that were potentially expressed by a specific treatment, only data belonging to MYC amplified cell lines were considered.
- Of the data belonging to the MYC amplified type, only those samples obtained by silencing gene BRG1 were considered in order to combine these data with the ones in platform GPL21185.

All 17 samples belonging to the GPL21185 platform, which were acquired with the suppression of gene BRG1 by targeting it with shRNA, were used instead.
Finally, data belonging to the two platforms were merged to form a unique dataset, paying attention to discard those genes that were not expressed in both subsets.
A final note regards the fact that SCLC data belong to the H82 cell line while NSCLC belong to the H460 cell line.

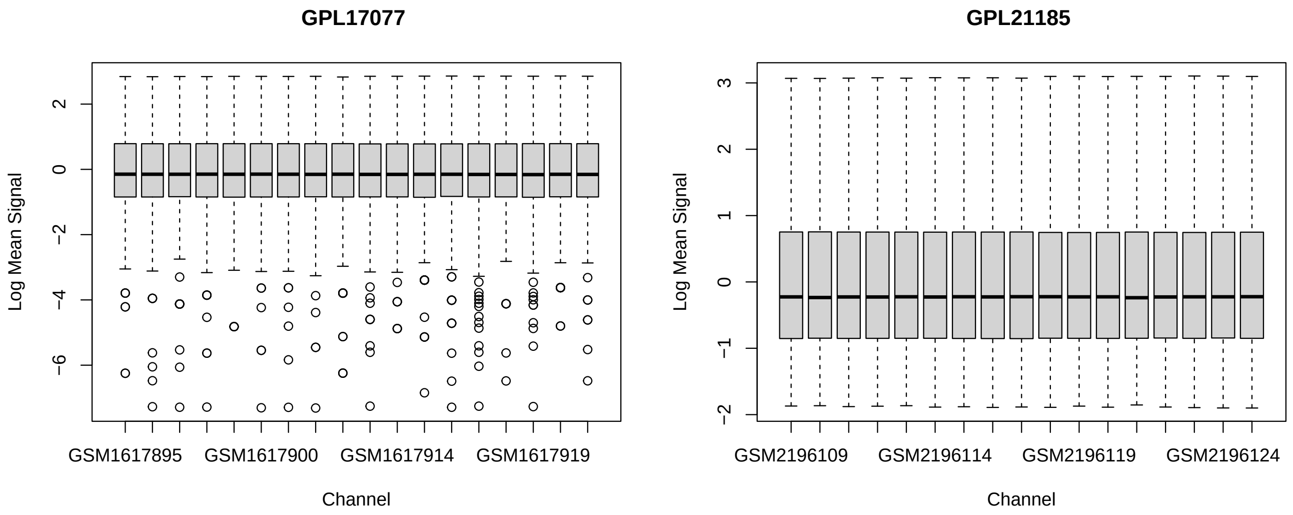
Exploratory analysis

3.1 Normalisation procedure

A normalisation procedure was applied to the whole dataset with the purpose of reducing noise and modifying data so that they could be easily compared. The normalisation consisted in two main steps:

1. To perform the logarithm base 2 of data
2. To perform a normalisation to a median of zero

The log-transformation was performed to make the distribution of data more symmetrical and thus more suitable for any statistical analysis. The second step involved both an alignment of boxes to a reference value of zero for the median and a rescale of the dimension of boxes to make them homogeneous.



(a) Boxplot of the pre-processed dataset GSE66245-GPL17077 after performing the normalization procedure.

(b) Boxplot of the pre-processed dataset GSE66245-GPL21185 after performing the normalization procedure.

Figure 1. Boxplot of pre-processed data obtained after performing both the logarithm base 2 of data and a median alignment.

A boxplot of both the pre-processed dataset GSE66245-GPL17077 (Fig.1a) and GSE66245-GPL21185 (Fig.1b) shows that in the normalisation procedure, the various error sources were discarded. This was a guarantee that, even though data belonged to two different platforms, the whole dataset did not present relevant differences due to artifacts or that could be attributed to other biological reasons.

3.2 Principal component analysis : PCA

Principal component analysis (PCA) is a technique used to analyse large amount of data to reduce the dimensionality of the original data set. PCA can be interpreted as a linear change of basis that highlights directions of maximum variance and projects data onto these new axes, called *principal components* (PCs). In this work, PCA was applied to gene expression. Even though PCs lack any physical meaning, the purpose was to identify relevant directions that could highlight particular features of data that were hidden if projected onto the original directions.

Table 1. Percentage of explained variance by the first 3 PCs together with the percentage of cumulative explained variance for the two separated datasets and the final combined one.

Sample	Var(PC1)	Var(PC2)	Var(PC3)	Cumulative variance
GPL17077	0.58	0.06	0.05	0.70
GPL21185	0.71	0.06	0.06	0.84
GPL17077+GPL21185	0.54	0.11	0.04	0.69

Explained variance is the quantity that expresses how much variation in the dataset can be attributed to each of the principal components. By telling how much of the total variance is 'explained' by each component, explained variance allows one to rank components in order of importance: the larger the variance accounted for by the component, the more that component is important.

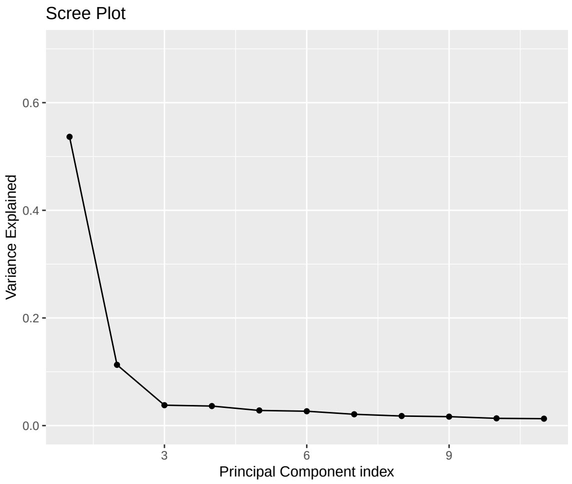


Figure 2. Scree plot of the combined dataset representing percentages of the variance explained by the first 11 PCs.

Results concerning the variance explained by the first three PCs together with the cumulative variance are reported in Table 1. Both the cumulative variance calculated for the first three components and the variance of PC1 were higher when datasets were considered separately compared to the joined one. This may suggest that the direction of PC1 was diverse in the two datasets. Thus, when data were merged, PC1 was not the optimal one for either of the two single datasets but it was probably a combination of their two principal axes.

The scree plot for the whole dataset is reported in Figure 2. As expected, the first three PCs account for the most data variability meaning that the largest amount of information that can be gathered from the dataset can be reached and visually inspected by considering a 3-dimensional space.

Projections onto the PC1-PC2 plane of the whole dataset (Fig.3) showed that PC1 was the direction that established a clear separation between genes that express SCLC cells (blue and yellow) from those expressing NSCLC cells (red and green). PC2 was the direction along which a separation between data belonging to platform GPL17077 (yellow and green) and those belonging to GPL21185 (blue and red) was detected.

PCA was then performed on expression profiles of the two platforms separately to understand the role played by the different types of treatment on both datasets.

Concerning platform GPL17077, PC2 was the direction from which more information about the treatment type could be gained. From Figure 4a it can be seen that FBS, for both SCLC and NSCLC, had the highest values of PC2 while more similar results were obtained for treated cells. Projections onto the PC1-PC3 plane of the dataset

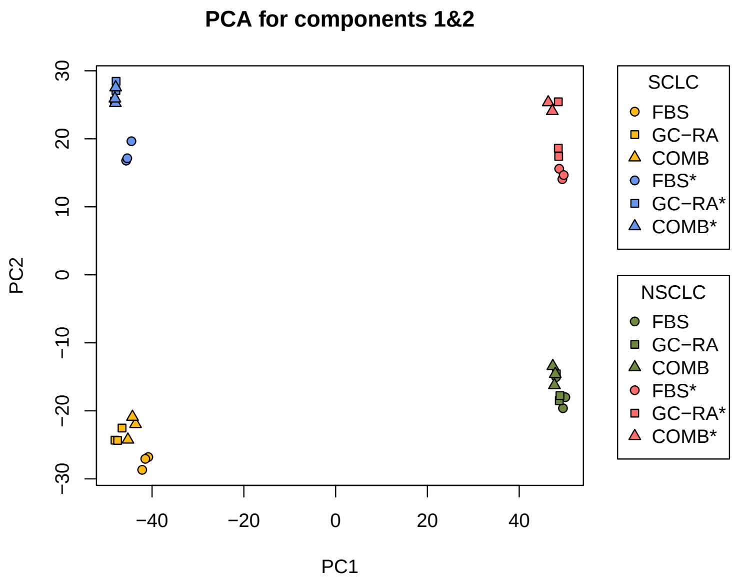
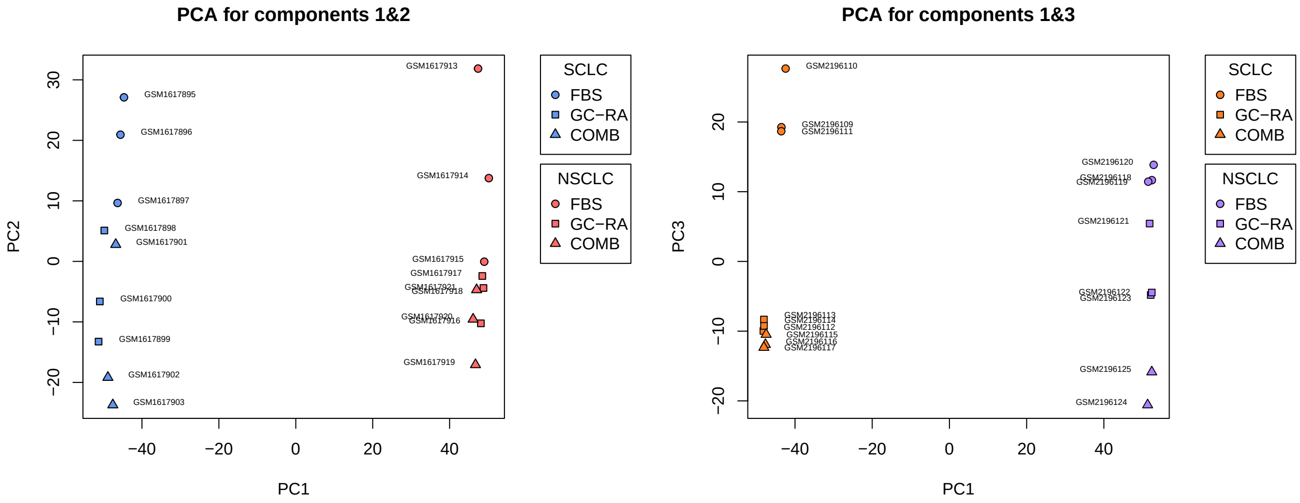


Figure 3. PCA of the whole dataset. The * in the legend distinguish data that belonged to platform GPL21185 from the ones from GPL17077. In the plot different colors are used to separate data both for type and platform.



(a) PCA plot of component 1 and 2 of pre-processed data originally from platform GPL17077.

(b) PCA plot of component 1 and 3 of pre-processed data originally from platform GPL21185.

Figure 4. PCA plots comparing gene expression for SCLC and NSCLC cancer cell types. Different treatments are highlighted with diverse shapes.

GPL21185 (Fig.4b) show that a clear separation between treatments can be observed along PC3 instead. Regarding SCLC data treated with both GC-RA alone and in combination with AZA/SAHA, a unique cluster was formed while non-treated data displayed a higher value of PC3. This means that along PC3 it was possible to distinguish between treated and non-treated data. Also in the case of NSCLC, non-treated data appeared to have the highest value of PC3 but, in this case, two distinct clusters were observed for patients treated with GC-RA alone and in combination with AZA/SAHA.

Unsupervised learning methods

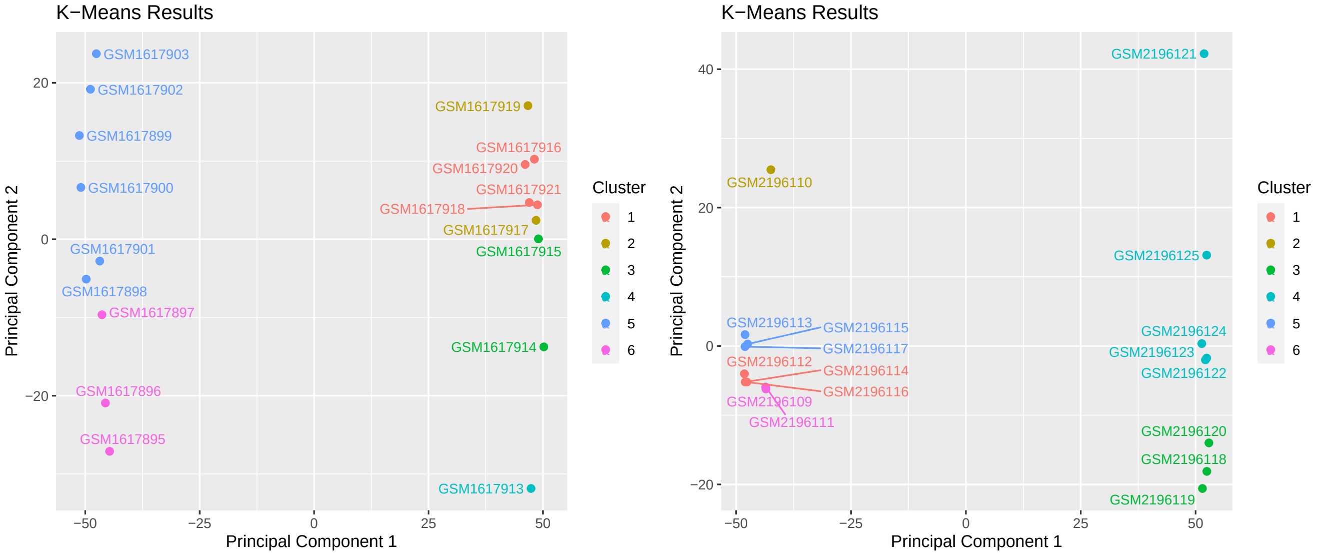
Data clustering is an approach often used to identify similarities between data points by grouping observables that display common features. Since clustering identifies a general task to be solved and not a specific classification procedure, different algorithms have been developed. The choice of the algorithm depends on the features of the selected dataset. Moreover, the clustering procedure is intrinsically subjective, mainly due to the freedom of choosing the type of linkage and method used to label different clusters. Hence, distinct results can be obtained starting from the same input data. In this project, two different clustering techniques were used to analyse data, namely the K-means algorithm and hierarchical clustering.

4.1 K-means algorithm

K-means belongs to the class of partitional clusering algorithms. A parameter k that represents the expected number of clusters must be given in input to the algorithm. Initially, a set of k random centroids $\{\mathbf{m_j}\}, j = 1, ...k$ is chosen and points $\mathbf{x}$ are assigned to the closest cluster C_j. Then, the algorithm proceeds iteratively by re-computing centroids measuring the Euclidean distance from all the points belonging to that group. Convergence is reached when the sum of square errors (SSE)

$$SSE = \sum_{j=1}^{k} \sum_{x \in C_j} \text{dist}(\mathbf{x}, \mathbf{m_j})^2 \qquad (1)$$

is minimized.
From the previous exploratory analysis, a number of 6 clusters corresponding to the FBS cells and the pair of treatments employed for both cancer cell types were expected to be observed in the data of both platforms. Similarly to PCA, data from the two platforms were considered separately to allow easier detection of the role played by different treatment types.



(a) K-means algorithm performed on pre-processed data originally from platform GPL17077.

(b) K-means algorithm performed on pre-processed data originally from platform GPL21185.

Figure 5. K-means plots comparing gene expression for SCLC and NSCLC cancer cell types. Clusters obtained from this procedure, corresponding to different treatment types, are identified by different colors.

Regarding samples from platform GPL17077, K-means clustering results (Fig.5a) were different compared to the ones reported for PCA (Fig.4a). A first consideration is that the algorithm was able to distinguish between the leftmost and rightmost groups, corresponding to the SCLC and NSCLC cell types respectively, since there were no points of the 6 detected clusters shared by the two major groups. Moreover, in the case of SCLC, the algorithm was able to correctly distinguish between cells grown in FBS and treated ones, even though a separation between the two different treatment types was not highlighted. In the case of NCSLC cell types clusters obtained from the K-means method did not correctly represent different treatment types.

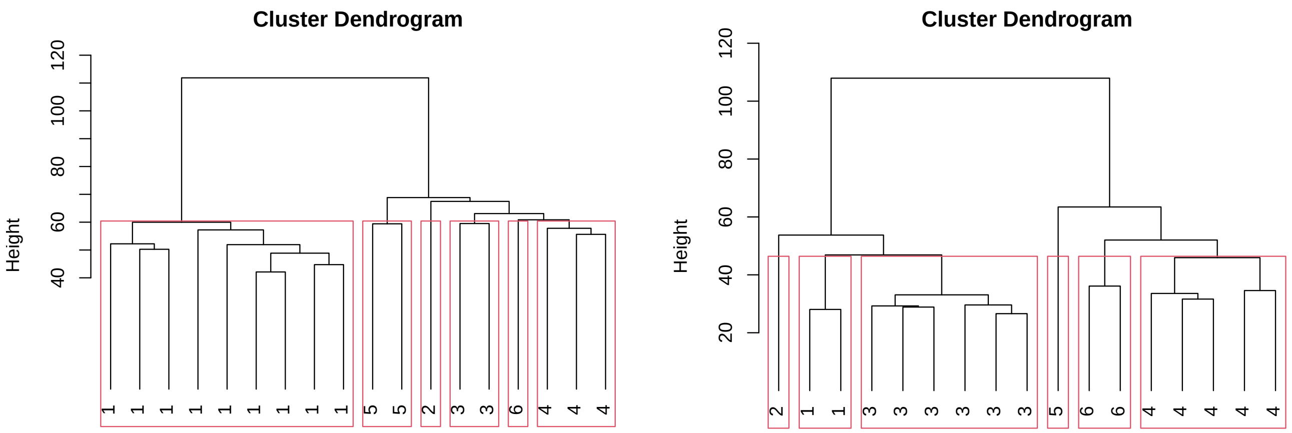
In the case of platform GPL21185 (Fig.5b), the K-means method displayed more difficulties in grouping samples correctly, based on the type of treatment. The only significant cluster corresponded to cells grown in FBS for the NSCLC cancer type.

In both datasets, the algorithm was not able to correctly classify the treatment type but it was eventually able to distinguish between cells that underwent a treatment from those which were simply grown in FBS. This could mean that the gene expression for cells in FBS was significantly different compared to the treated ones, which, on the contrary, did not show any feature that the algorithm could use to distinguish between treatment types. This observation may suggest that AZA/SAHA had a lower impact on the expression profile of genes compared to glucocorticoids and retinoic acid.

4.2 Hierarchical clustering

A second method used to perform cluster analysis on the dataset exploits a hierarchical bottom-up approach, according to which not only the distance between points (intra-cluster distance) was evaluated, but also an inter-cluster distance. An Euclidean distance metric was chosen to compute the dissimilarity between each observation in the dataset.

In this project the average-linkage method was exploited to compute the inter-cluster distance, which was established by calculating the average pairwise distance of all pairs of objects from different clusters. Subsequently, dendrograms (Fig.6) were plotted as a tree representation of the arrangement of groups obtained from the previous linkage choice. In this method, a distance threshold was fixed, and the number of clusters for the dataset resulted consequently. In both considered datasets six clusters were detected by the algorithm. This result was in agreement with the conclusions drawn from the K-means method.



(a) Dendrogram computed from pre-processed data originally from platform GPL17077.

(b) Dendrogram computed from pre-preocessed data originally from platform GPL21185.

Figure 6. Dendrogram plots resulting from the hierarchical clustering algorithm.

Supervised learning methods

A second possible approach to analyse data exploits supervised learning methods, in which labels are associated with data points. The aim of this approach is to build a function capable of predicting a classification of data, based on what the algorithm learns. Supervised learning methods are often based on two major phases :

1. *Training phase:* the algorithm learns features of input data based on their labelling
2. *Validation phase:* the algorithm figures out an accurate classification of new data points, based on the results of the training part.

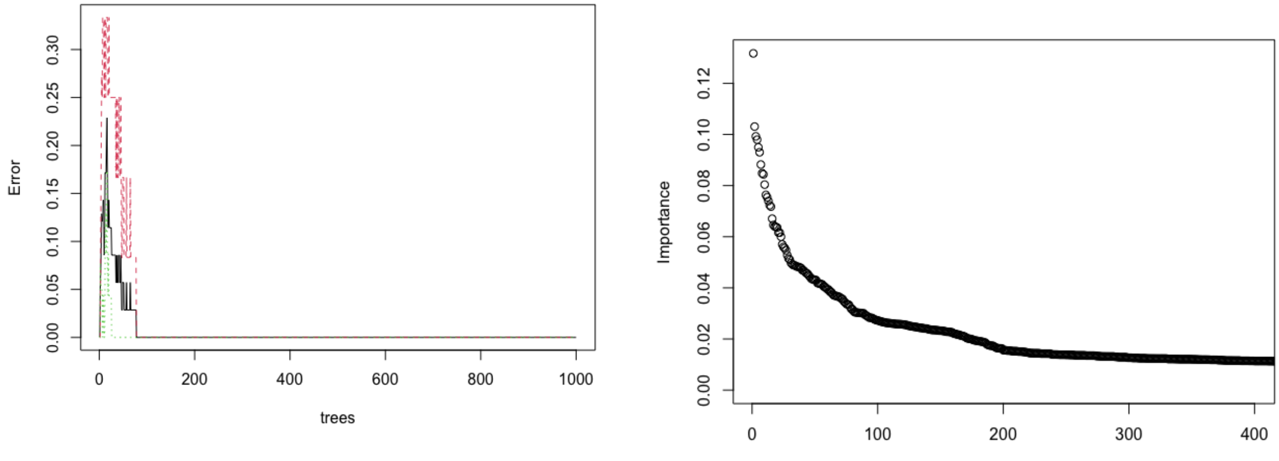
The aim of this analysis was to compare various classification methods to determine the one that could best sort data points and extract the most relevant genes to be further investigated through functional analysis.

It is worth mentioning that from the previous exploratory analysis, it emerged that not only a grouping based on the cancer type (SCLC and NSCLC) but also a sorting based on the type of treatment could be exploited on expression profiles. However, since the focus of this work was on the detection of differences in expression patterns of genes subject to treatment rather than detecting specific features proper of the type of cancer, only the second grouping approach was considered. Thus, samples were distinguished between *treated cells*, labelled as GCRA-COMB and *non-treated cells*, labelled as FBS.

5.1 Random forest

Random forest is an ensemble learning method which, for the purpose of classification, is based on the construction of multiple decision trees and whose outcome is the class selected by the majority of these.

The R function `randomForest()` was used to generate a forest of 1000 decision trees by first creating a bootstrapped dataset from the original one and then randomly selecting a subset of attributes at each step.



(a) Plot of the error in classification. The red line represents the error in classifying for FBS; green line is the error in classifying for GCRA-COMB; black line is the average.

(b) Plot of the importance ad a function of the gene index.

Figure 7. Plot of the importance as a function of the gene index and error in classification committed in the random forest algorithm.

In Figure 7a, the overall error computed by the algorithm as a function of the number of decision trees within the forest shows that perfect classification was reached using less than 200 trees. Moreover, the higher error computed in the classification of genes within the GCRA-COMB class (red dashed line) compared to the classification in the FBS one (green dashed line) suggests that for the algorithm it was more difficult to detect genes subject to treatment.

The *importance* variable, which identifies the main features in the classification by looking at the rate of appearance of each variable in trees, was then computed to rank the most relevant genes. The majority of genes ranked in order of importance (Fig.7b) displayed values lower than 0.12, suggesting that few of them were the most relevant ones. Thus, to reduce noise that could be generated by considering non-relevant terms in the analysis, a subset of 100 genes was selected to be further analysed from a more biological point of view.

A first glimpse into the biological relevance of these genes was obtained by computing a heatmap with the 25 most relevant genes (Fig. 8). The gene relevance is quantified through the Z-score, a statistical measure that represents the number of standard deviations by which the value of a data point is above or below the mean value of the whole observation. From the plot, the majority of the 25 most relevant genes belong to the class of samples treated using either GC/RA alone or combined with AZA/SAHA. 10-fold cross-validation was performed and the quality of classification computed with Random Forest was inferred. Results, quantified in terms of accuracy (Fig. 11a), were compared with the ones obtained from the other classification methods and are discussed in greater detail in the next sections.

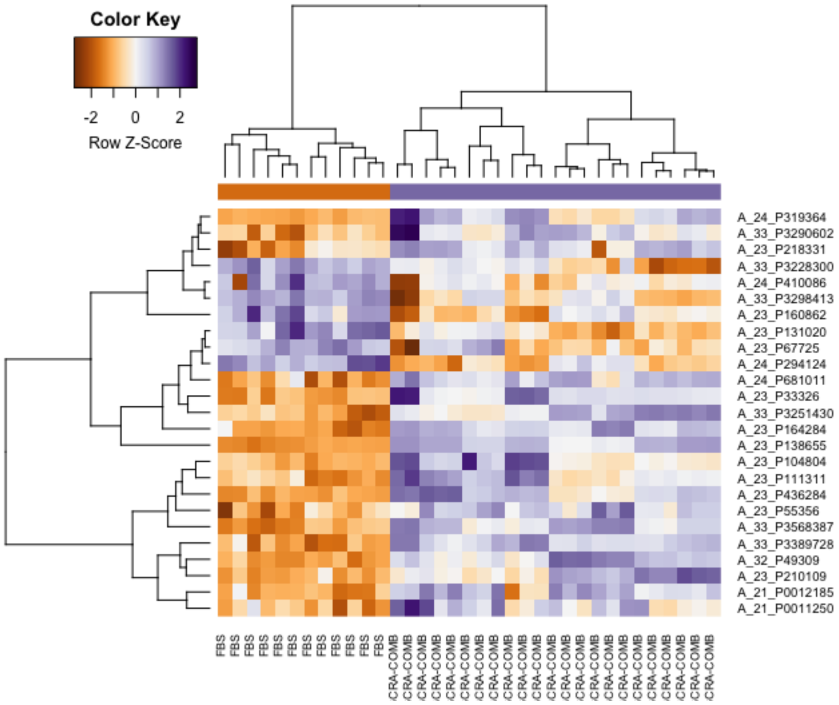
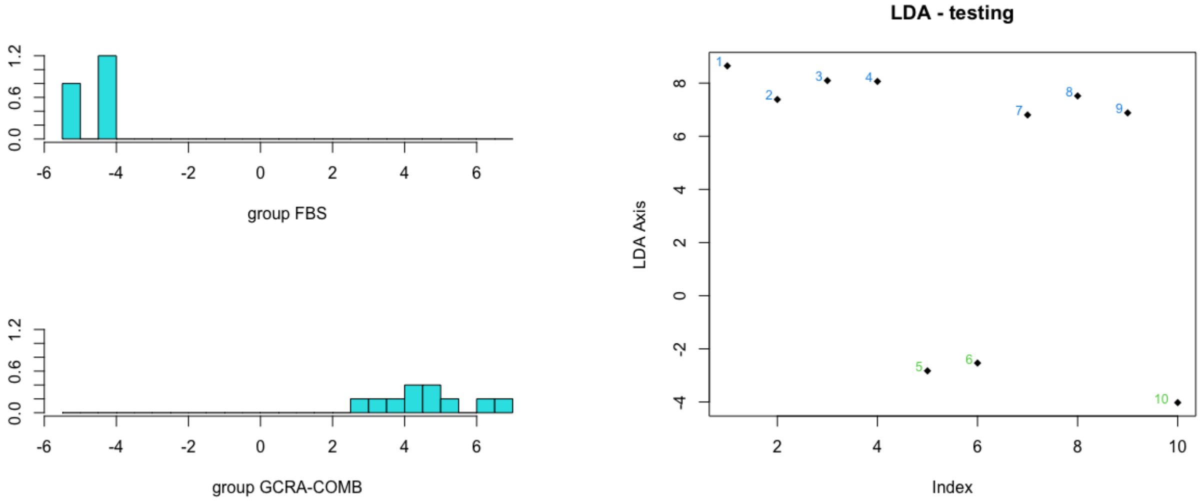


Figure 8. Heatmap showing the 25 most relevant genes. genes are named with the associated Agilent probe ID.

5.2 Linear discriminant analysis (LDA)

Linear discriminant analysis is another supervised learning method used in this work with the purpose of classification. This method is based on the idea of finding the 1D direction that maximizes the separation of data point projections into distinct groups.



(a) Plot projections of data points along the direction that maximizes the separation of the two groups.

(b) Plot of the classification of the testing dataset performed with LDA.

Figure 9. Plots resulting from the linear discriminant analysis performed on the whole dataset.

Also in this case, the method relies on the concept of distance. *Between-class distance* is defined as the distance between centroids of two separated groups while *within-class distance* represents the sum of distances of all data points from the centroid in each group. The direction that works the best for classification is one that both maximizes the between-class distance and minimizes the within-class distance.

Firstly, a row t-test was performed on the whole dataset thanks to which only genes whose p-value was lower than the threshold of 0.1 were kept for the analysis.

The whole dataset was then split into two groups:

- Training group : 15 samples from GSE66245-GPL17077
- Testing group : 10 samples from GSE66245-GPL21185

and the training subset was given in input to the `lda()` function. Projections of data points along the direction that maximized the separation between the two groups (Fig.9a) appear to be well divided. When LDA was used to predict the classification of the testing group (Fig. 9b), a perfect separation between the group of FBS (green) and GCRA-COMB (blue) was observed. This result is in agreement with the perfect classification of data points in both FBS and GCRA-COMB classes performed by the Random Forest algorithm.

A more robust separation between training and testing datasets was performed by exploiting functionalities of the `caret` package. In this case, a 10-fold cross-validation using accuracy as a metric was performed considering the previously defined training and testing datasets. Also in this case, a perfect classification of samples within the testing group was reached.
The quality of classification was then quantified in terms of *accuracy*. Figure 11a confirms that perfect classification was reached by exploiting both the Random Forest and LDA algorithms.

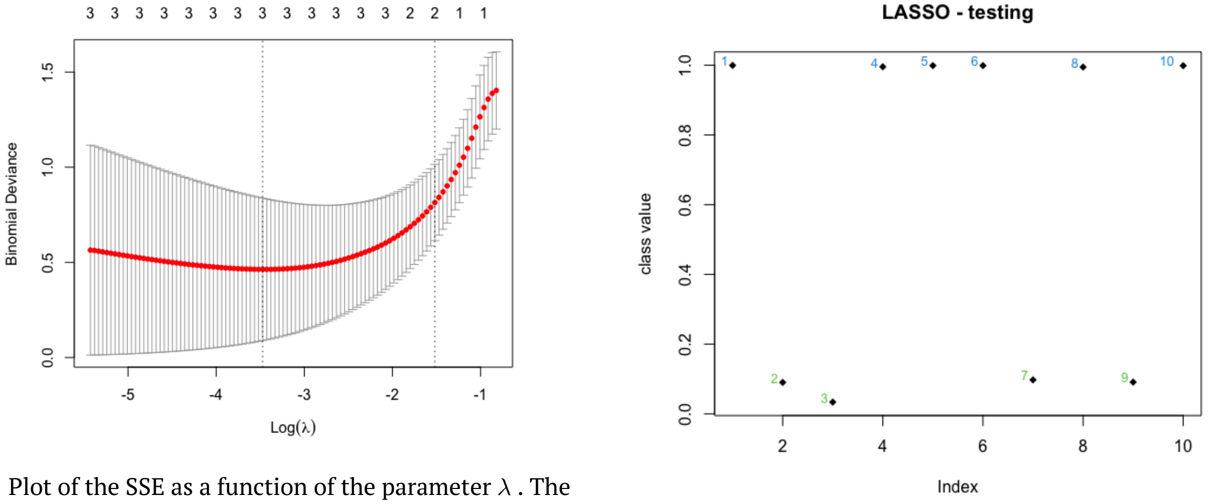
5.3 Least Absolute Shrinkage and Selection Operator (LASSO)

Linear regression is a way to model an association between variables which assumes a specific linear relationship between predictors $X = \{x_i\}$ and responses $Y = \{y_i\}$ according to

$$Y = \beta_0 + \beta_1 \cdot X \tag{2}$$

which could also be generalised for higher dimensional spaces. The aim of classical linear regression is to determine the unknown parameters β_i by minimizing the sum of square errors (residual sum of squares for higher dimension) between the linear model and data points. However, this estimate is often not reliable mostly because of the bias introduced by non-relevant data.

Least absolute shrinkage and selection operator (LASSO) is a regression analysis, originally developed for linear regression models, introduced with the aim of improving the prediction accuracy and the interpretability of regressions.



(a) Plot of the SSE as a function of the parameter λ . The leftmost dashed lines indicates the minimum of the curve while the right-most one indicates the value of λ that is separated by one std deviation from the previous one.

(b) Plot of the classification of the testing dataset performed with LASSO.

Figure 10. Plots resulting from the LASSO analysis performed on the testing subsets.

Through this analysis a feature selection was performed, according to which a subset of predictors relevant for the regression model was identified. This method implemented variable selection in terms of shrinkage of the number of β-coefficients that could be relevant for the linear model. This was achieved by introducing a weighted penalty in the SSE that had to be minimized:

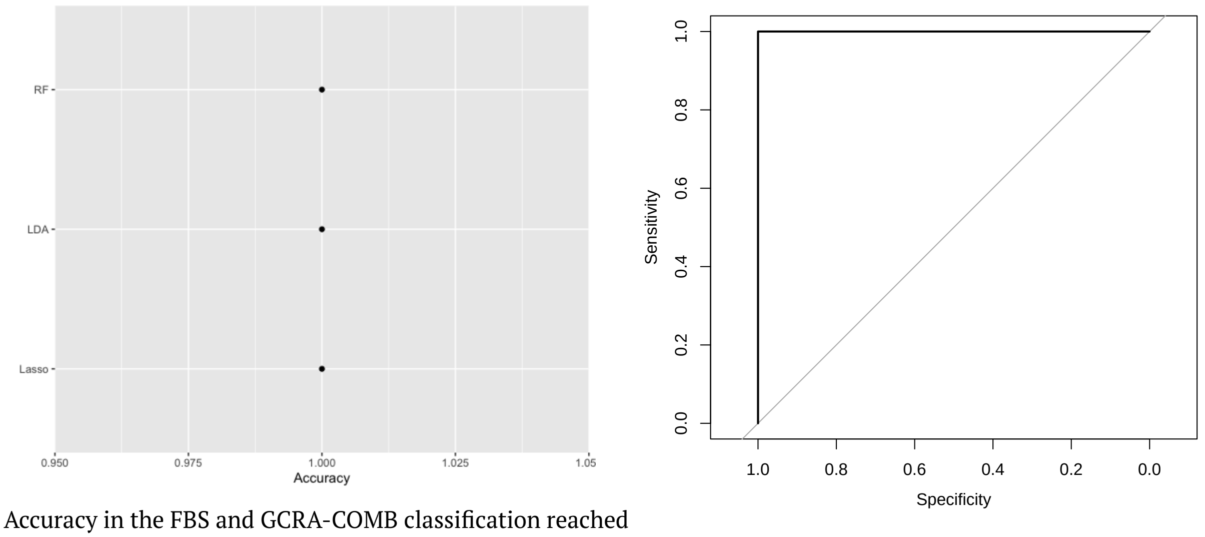
$$SSE(\beta) = \sum_{i}(y_i - (\beta_0 + \beta_1 \cdot x_i))^2 + \lambda \sum_{j} |\beta_j| \tag{3}$$

where $\lambda \in [0, 1]$ determined the strength of the shrinkage.

In this work, a linear relationship was assumed between the value of expression of genes (predictor) and labels (response). With the purpose of classification, the dataset was manually separated into a training (15 samples from GPL17077) and a testing part (10 samples from GPL21185).

The linear fit with penalized maximum likelihood was built using the function `glmnet()` where `family='binomial'` was chosen to force the logistic regression. To have a more robust estimate of the penalty, internal cross-validation was performed through the function `cv.glmnet()`. Results of the SSE as a function of the penalty reported that the estimate of λ which produced the smallest error was $\lambda = 0.03$. Once the model was built from the whole dataset, training was performed and predictions were made on the testing subset.

As it can be seen in Figure 10b, classification between the two classes (0 : FBS and 1: GCRA-COMB) was perfectly reached also with this method. A more robust classification was also attempted exploiting the `caret` package, as in the methods previously illustrated. The quality of the classification was assessed by performing a 10-fold cross-validation, from which the best value of the penalty was estimated to be $\lambda = 0.25$, as reported in Figure 10a. With this additional step, unitary accuracy was estimated also for this method (Fig. 11a).



(a) Accuracy in the FBS and GCRA-COMB classification reached using the three superivised methods combined with a 10-fold cross validation: random forest (RF), LDA and LASSO.

(b) Plot of the ROC curve obtained for Random Forest, LDA and LASSO.

Figure 11. Plots of the performances of classification methods.

The performance was also visually represented through the *Receiver Operating Characteristic* (ROC) curve and an overall assessment of the prediction quality was obtained by calculating the *area under the curve* (AUC). As it can be seen in Figure 11b, since perfect classification was obtained, the ROC curve corresponded to the ideal one and the AUC was equal to 1. Despite this result, the recognition of the reliability of results obtained from Random Forest among the scientific community led to choose this supervised learning procedure to extract the most relevant genes.

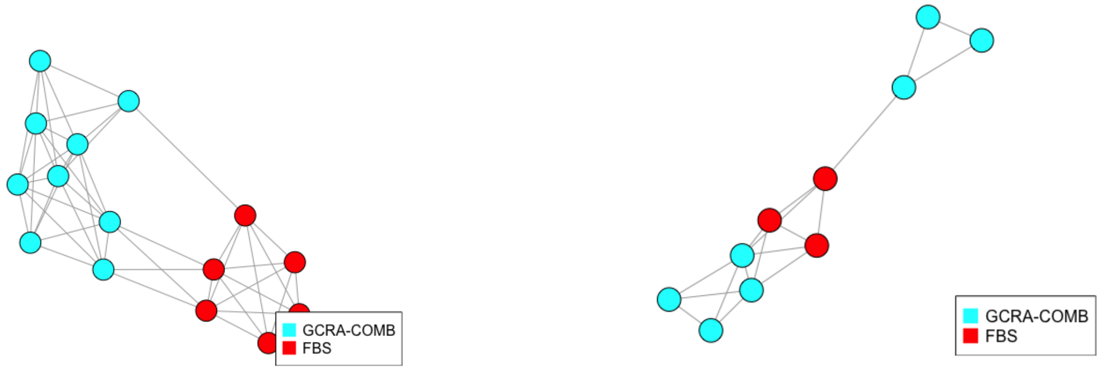
5.4 SCUDO classification

SCUDO (Signature-based Clustering for Diagnostic Purposes) is a rank-based method for the analysis of gene expression profiles developed by COSBI[1] and used with the purposes of classification and diagnosis [2].

A set of gene signatures, namely a group with the 50 most up-regulated and 50 down-regulated genes, was identified for each sample in the training dataset exploiting the function `scudoTrain()` included in the `rScudo`

[1] The Microsoft Research University of Trento Centre for Computational and Systems Biology

package. This function also implemented an all-to-all comparison between signatures, and a distance matrix, which quantified the amount of similarity between expression patterns of different samples, was built. Validation was then performed on the testing dataset through the function `scudoTest()`, exploiting features selected in the training phase. A graph of samples, computed through the function `scudoNetwork()`, allowed to express which samples were characterized by a level of similarity, quantified through GSEA, greater than a fixed threshold `N=0.4`. The connected graph obtained from the training dataset (Fig.12a) shows that the algorithm could establish a clear separation of samples belonging to the FBS group from the ones in GCRA-COMB. However, in the network of samples generated from the testing set (Fig.12b) a clear separation in two groups was not observed.



(a) Graph of samples computed on the training dataset (15 samples of GPL17077) using the SCUDO algorithm.

(b) Graph of samples computed on the testing dataset (10 samples of GPL21185) using the SCUDO algorithm.

Figure 12. Graph of samples computed exploiting the SCUDO algorithm

The function `scudoClassify()` was finally used to perform a supervised classification of test samples exploiting the model built in the training phase. Performances were quantified using the `caret` package obtaining as a result a unitary accuracy, quantifying the perfect classification of the 3 samples of the FBS group and the 7 samples in the GCRA-COMB one.

Functional Enrichment Analysis

Functional enrichment analysis is a bioinformatic procedure which aims at discovering biological annotations that are over-represented in a list of genes with respect to a reference background. This procedure is often used to identify biological pathways/processes that are particularly abundant in a list of genes. Biological pathways describe molecular activities and identify the role played by genes in performing certain biological functions. Enrichment analysis was computed on the subset of the 100 most relevant genes ranked by importance value obtained from the Random Forest algorithm, exploiting the functionality g:GOSt of the web-based tool g:Profiler [3].

Table 2. Results of the functional enrichment analysis obtained exploiting the g:Profiler web based tool.

Source	Term ID	Term name	p-value
GO-BP	GO:0048387	negative regulation of RA receptor signaling pathway	7.417×10^{-3}
GO-BP	GO:0042573	retinoic acid metabolic process	1.325×10^{-2}
GO-MF	GO:0008401	retinoic acid 4-hydroxylase activity	3.706×10^{-3}
KEGG	hsa:00830	retinol metabolism	2.026×10^{-2}
WP	WP:716	vitamin A and carotenoid metabolism	1.041×10^{-2}
HPA	HPA:0461392	skin 1 : cells in basal layer	4.94×10^{-2}

P-values, quantifying the enrichment of a pathway/process, were computed using the Fisher's exact test and the default g:SCS algorithm was chosen to correct for multiple-testing. Individual terms beyond a significance threshold of 0.05 were deemed enriched. All other parameters in the tool were left to their default value.

The most enriched GO term related to the biological process suggests a down-regulation of genes coding for the retinoic acid receptor (RAR), a type of nuclear receptor that is often activated by all-trans retinoic acid (ATRA) and found within the nucleus of cells. Suppression of retinoic acid receptor was investigated in different studies, some of which concerned biological implications in lung cancer. According to Xu *et al*, the use of retinoids in some patients with NSCLC was able to prevent second primary tumours. However, as soon as some types of RARs (mostly RARβ) and retinoid X receptors (RXRs) were suppressed, an abnormal activity that could enhance cancer development was observed [4].

A deeper investigation of the role played by RA in the treatment of lung cancer was performed combining results obtained from functional enrichment analysis with *network-based analysis*, a method that accounts for interactions among functionally related genes. Three tools, based on different approaches, were used:

- the web-based tool STRING [5];
- the web-based tool EnrichNet [6];
- the `R` package `PathfindR` [7].

6.1 STRING

STRING is a software used to detect subgroups of correlated genes among a provided list of genes. It is based on the identification of both physical and functional protein-protein interactions, generated by exploiting information built from prior knowledge gained from sources such as databases or genomic context predictions.

A list of proteins, encoded by the 100 most relevant genes ranked by importance through the Random Forest algorithm, was provided to the tool. However, the protein-protein interaction network (PIN) displays the highest level of confidence[2] was only made by a subset of 7 biologically correlated proteins, among which 3 belong to the cytochrome P450 family. These enzymes play an important role in the oxidization of substances and are involved in reactions such as hydroxylation.

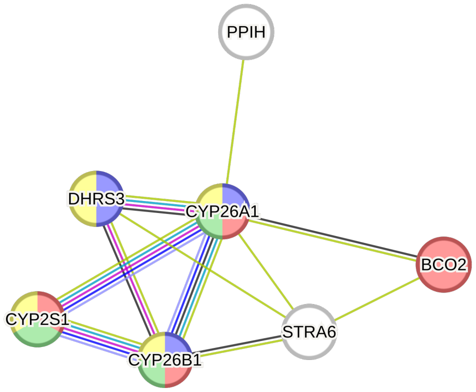


Figure 13. Protein-protein interaction network on 7 most relevant genes obtained performing a network-based analysis exploiting the web-based tool STRING. Color code GO:0042573, GO:0048387, GO:0008401, hsa:00830.

The network reported in Figure 13 shows that the majority of known interactions, both curated from databases and experimentally observed, involve proteins of the cytochrome P450 family and the short-chain dehydrogenase/reductase 3, encoded by the DHRS3 gene. The latter enzyme catalyzes the oxidation/reduction of a wide range of substrates, including retinoids. The colour code identifies pathways/processes to which genes belong to:

- CYP26A1 and CYP26B1 are involved in the enriched KEGG pathway and all processes detected with functional enrichment analysis;
- CYP2S1 is involved in the KEGG pathway and both the retinoic acid metabolic process and retinoic acid 4-hydroxylase activity;

[2] Here confidence is considered as the likelihood of STRING judging an interaction to be true.

- DHRS3 is involved in the KEGG pathway and negative regulation of RA receptor signaling pathway;
- BCO2 is only involved in the retinoic acid metabolic process;

As it can be seen in Table 13, correlations with the highest confidence scores[3] (> 0.9) involve gene CYP26A1 interacting with DHRS3 and with CYP2S1, and gene CYP26B1 interacting with CYP2S1.

Table 3. Combined scores quantifying the level of confidence of interactions among the subset of 7 genes selected for the network-based analysis using STRING. Only interactions with a confidence score > 0.5 are reported.

ID gene 1	ID gene 2	Combined score
DHRS3	CYP26A1	0.968
CYP26A1	CYP2S1	0.935
CYP26B1	CYP2S1	0.933
CYP26B1	CYP26A1	0.827
CYP26A1	STRA6	0.771
CYP26B1	STRA6	0.678
CYP26A1	PPIH	0.648
DHRS3	CYP26B1	0.542
BCO2	STRA6	0.530

6.2 EnrichNet

The assessment of functional relations among genes/proteins in biological processes can be improved exploiting the web-base application EnrichNet. Differently from other tools, EnrichNet complements the basic enrichment analysis with a distance measure that quantifies how close the list of genes given in input is to a set of genes related to a specific function. The relevance of gene ontology terms and pathways is quantified in terms of a network-based association score (the XD-score), which measure the network interconnectivity between the set of genes/proteins in input and the cellular pathways/processes mapped to the molecular interaction network.

Few biologically relevant processes and no biological pathways (Tab. 4) were obtained exploiting this tool on the set of 100 most important genes given in input. No additional information was gained from the provided network visualization.

Table 4. Ranking of GO terms obtained exploiting the EnrichNet web based tool based on XD-score.

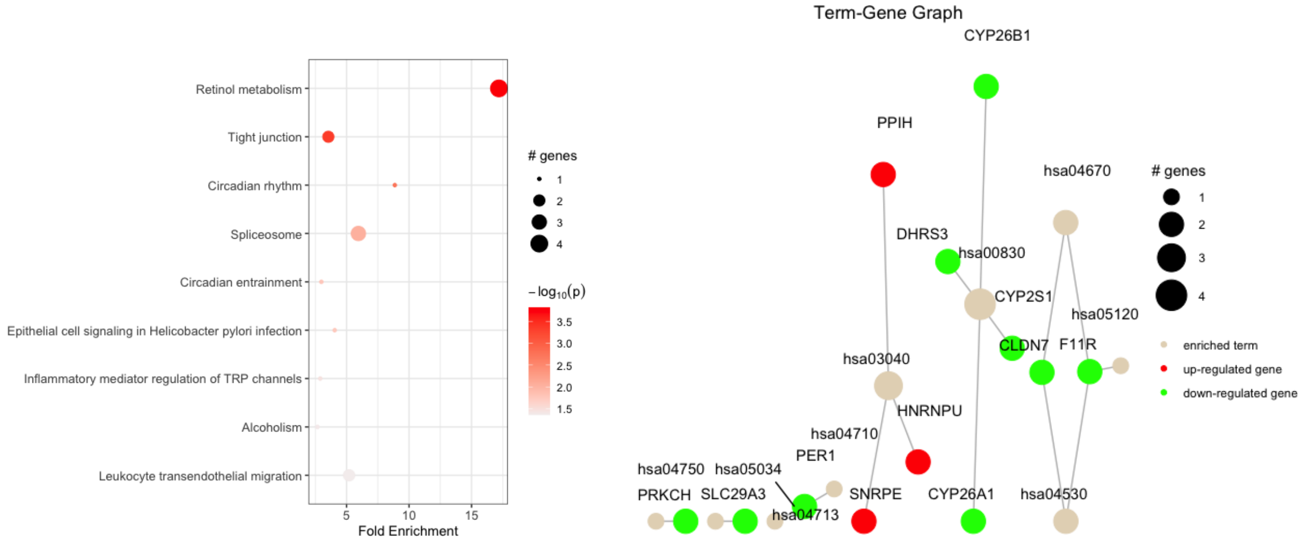
Source	Term ID	Term name	XD-score	Overlapping genes
GO-BP	GO:0045070	positive regulation of viral genome replication	1.8	NR5A2, TARBP2
GO-BP	GO:0001523	retinoic acid metabolic process	1.2	BCO2, CYP26A1
GO-BP	GO:0048384	retinoic acid receptor signaling pathway	1.0	CYP26A1, CYP26B1

6.3 PathfindR

PathfindR is a R package used to perform an *active-subnetwork-oriented enrichment analysis*, according to which basic enrichment analysis is joined with the identification of active subnetworks, namely groups of interconnected genes in PIN that correspond to significantly enriched or depleted ones. This analysis was performed through the wrapper function `run_pathfindR()` to which the list of 100 most significant genes was given in input together with the associated p-values calculated performing a row t-test.

Enrichment results are displayed in terms of a dotplot (Fig.14a) in which the 10 most significant KEGG pathways are reported. The retinol metabolism (hsa:00830) resulted to be the most significantly enriched pathway, in agreement with results obtained from functional enrichment analysis exploiting g:Profiler. Additional information was gained by plotting the significant genes involved in the enrichment terms, as shown in Figure 14b. Concerning the retinol metabolism, CYP26A1, CYP26B1, CYP2S1 and DHRS3 were the most up-regulated genes, meaning that they were more expressed in the GCRA-COMB cells compared to the FBS ones.

[3] Indicates the approximate probability that a predicted link exists between two proteins in the same metabolic map.



(a) Dotplot of the 10 most significant pathways KEGG pathways.

(b) Plot with the significant up-regulated and down-regulated genes in the most enriched pathways.

Figure 14. Plot of the most enriched terms together with the associated significantly activated genes obtained exploiting the pathfindR tool.

The up-regulation of these genes, in particular of CYP26A1 and CYP26B1, has a fundamental implication to explain the role played by RA in treating tumors. CYP26A1 has biological relevance since it rules the cellular level of retinoic acid involved in the regulations of gene expression in tissues. CYP26B1 is a critical regulator of ATRA levels by inactivating it into the hydrolyzed form instead. The up-regulation of these genes in GCRA-COMB cells, causes a significant level of retinoic acid to be inactivated. Lower levels of ATRA trigger the suppression of RAR, thus causing the possible enhancement of cancer development suggested by Xu *et al* [4]. The anti-tumoral nature of RA is related to its ability to induce cellular differentiation which, however, has a limited clinical efficiency against different types of tumours. Accordingly, studies carried out to determine the efficiency of retinoids in the specific cure of lung cancer did not display any clear benefit in either SCLC or NSCLC [8]. Results of this work, which could explain the resistance of cancer to a treatment based on retinoic acid, are also in agreement with studies carried out by Shelton *et al* in which the authors have observed an increased expression of CYP26A1 as a consequence of ATRA degradation

Methylation profiles

Partial Least Square regression is an alternative to ordinary least squares for handling multiple collinear data, such as in the analysis of high throughput biological data. Given two datasets, this method aims at finding a set of latent variables such that:

1. best explain for the X-space
2. best explain for the Y-space
3. express the greatest relationship between the X-space and Y-space

Latent variables are constructed as a linear combination of the original variables, weighted by coefficients that are referred to as loadings. These new directions are found by maximizing the covariance between projections of observables in the X-space and Y-space (scores) along these directions. The sparse version of PLS regression (sPLS) performs both a variable selection on datasets and seeks a linear combination of variables in order to reduce the overall dimensionality of the space.

In this work, sPLS was performed on a dataset created by joining a part of the expression profile (X) dataset with the methylation (Y) one, contained in the GPL13534 platform. This analysis aimed at revealing the presence of a relationship between the expression and methylation profile, identifying those genes that were more likely to be highly transcribed or repressed.

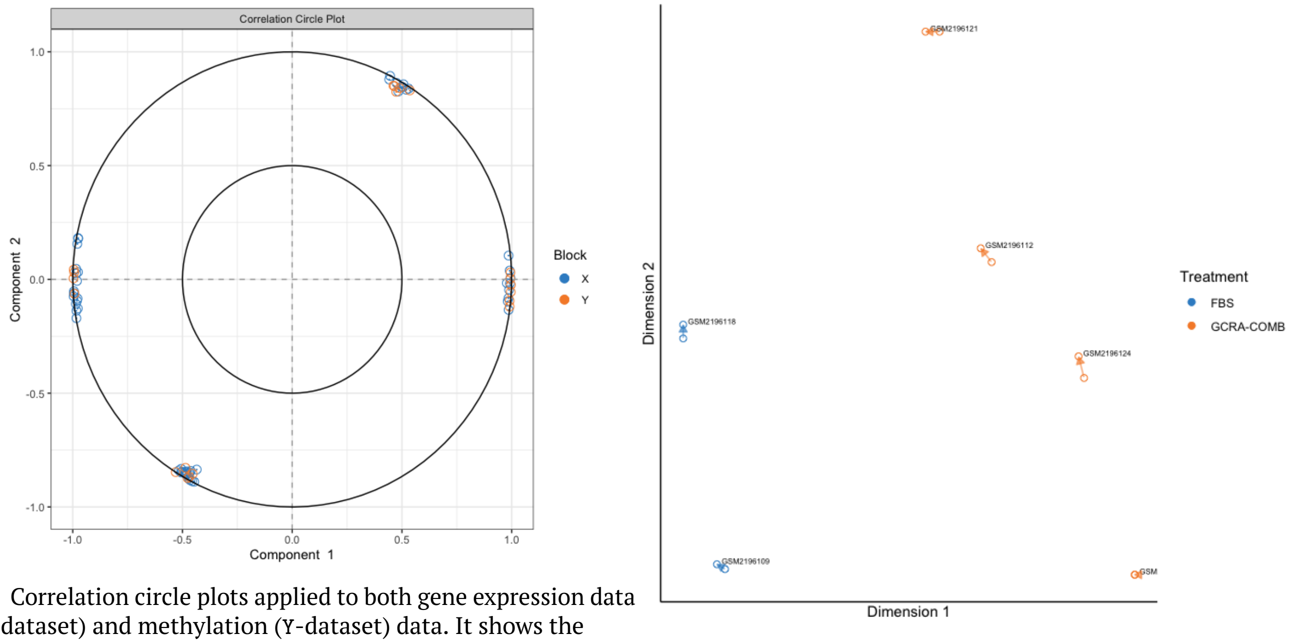
Firstly, the dataset was prepared considering

- for the *expression profile*: the first sample per treatment type among the ones previously selected in the GPL21185 platform;
- for the *methylation profile*: samples of SCLC and NSCLC belonging to the MYC amplified cell lines, with silenced gene BRG1;

Attention was paid to referring each sample of the methylation profile with the one belonging to the same patient in the expression profile. A row t-test was performed on both datasets, to select statistically significant data corresponding to a p-value lower than the threshold fixed at 0.1.

A model with 2 components was built exploiting the function `spls()` included in the `R` package `mixOmics`. A number of 25 variables for the expression profile dataset and 10 variables for the methylation profile were selected for each latent component.

Various information concerning correlation can be gained by exploiting the function `plotVar()`, in which variables that are kept to build the model are represented as points corresponding to vectors centred around the origin and normalized to 1. Results in Figure 15a show that a strong correlation exists between those genes and the resulting two latent variables since almost all points lie on the unitary circumference. From the plot, it can be seen that those genes kept to build component 1 almost do not count for component 2 since they contribute with a negligible weight to the latter component. Genes selected to build the second latent variable have a not negligible contribution towards both components, thus meaning that they display a strong correlation to both of them. Genes and chromosome locations are displayed as blue and orange points respectively in the plot. No names are reported both to avoid it being too crowded and because, among the selected genes, no one displayed a relevant role in the previous analysis.



(a) Correlation circle plots applied to both gene expression data (`X`-dataset) and methylation (`Y`-dataset) data. It shows the correlation structure in the space spanned by latent component 1 and 2. Variables were not named to avoid reporting a plot that is too crowded.

(b) Arrow plot representing samples projected onto the first two latent components. Samples were named according to the sample name present in the expression profile dataset

Figure 15. Correlation plots performed on the expression and methylation datasets showing both variables and samples.

Not only a correlation to latent components but also a relationship between variables of the two datasets can be obtained from this plot. Within the same dataset, genes selected for the two components are highly correlated since they appear to be very close to each other. An overlap is also observed among variables belonging to the two datasets for both components, thus meaning that a correlation between expression and methylation profiles exists.

Agreement between pairs of points in the expression and methylation datasets is reported in terms of arrow plot (Fig.15b). The presence of only short arrows indicates that a good agreement level exists between all samples.

Conclusions

In this work the role played by glucocorticoids and retinoic acid, both alone and in combination with AZA/SAHA was investigated by analyzing both the expression and methylation profile of a pre-processed dataset concerning lung cancer.

Unsupervised learning methods were able to distinguish between the SCLC and NSCLC groups but they were unable to correctly separate treated cells (FBS) from the non-treated ones (GC-RA and COMB).

Perfect classification of samples, quantified in terms of unitary accuracy, between treated and non-treated classes was reached exploiting supervised learning methods, through a 10-fold cross validation. The set of 100 most significant genes was extracted through the Random Forest algorithm, whose outcomes are well recognized by the scientific community.

Functional enrichment analysis performed exploiting g:Profiler highlighted the down-regulation of genes coding for RAR, which could trigger for cancer development, as suggested by Xu *et al*. Deeper insights were gained combining these results with Network Based Analysis. Genes CYP26A1 and CYP26B1, which are involved in the inactivation of ATRA, appeared to be the most up-regulated ones in the class of treated cells, thus triggering the inactivation of RAR. Since the suppression of this receptor could enhance cancer development, the obtained results were in agreement with the work of Hunsu *et al* according to which retinoids did not display any clear benefit in the treatment of lung cancer.

References

1. Steimbach, R. *et al.* Aza-SAHA Derivatives are Selective Histone Deacetylase 10 Chemical Probes That Inhibit Polyamine Deacetylation. *ChemRxiv.* (2021).
2. Lauria, M. Rank-based transcriptional signatures. *Systems Biomedicine* **1,** 228–239 (2013).
3. *gProfiler* https://biit.cs.ut.ee/gprofiler/gost.
4. Xu, X. C. *et al.* Suppression of Retinoic Acid Receptor in Non-Small-Cell Lung Cancer In Vivo: Implications for Lung Cancer Development. *JNCI: Journal of the National Cancer Institute* **89,** 624–629 (1997).
5. *STRING* https://string-db.org.
6. *enrichNet* http://www.enrichnet.org.
7. *pathfindR* https://egeulgen.github.io/pathfindR/index.html.
8. Hunsu, V., Facey, C., Fields, J. & Boman, B. Retinoids as Chemo-Preventive and Molecular-Targeted Anti-Cancer Therapies. *Int J Mol Sci.* **14** (2021).